Data Science Capstone Project

By Marc Hauptmann

A comparison of Chicago neighborhoods in terms of business and social activity against crime rate.

Introduction / Business Problem:

Crimes rates in Chicago are higher than the US average, especially when it comes to violent crimes such as murder and rape [Ref1]. Various studies have been conducted and articles written about the potential causes of the peak in crime rate in the city in recent times (e.g. [Ref2]). Potential causes such a decline in educational or social spending or the reduction of police presence and activity on the streets, have been widely dismissed. Underlying social-economic causes are more challenging to assess, due to the complex interactions and / or only incomplete mapping of the underlying contributing factors.

This Data Science Project is an attempt at solving 2 questions of importance to the 'bigger' problem described above:

- Is there a relationship between types of business venue activity in Chicago neighborhoods and the frequency and type of crimes committed?
- How do neighborhoods in Chicago compare in terms of 'prosperity' and 'safety' (i.o.w. what are safe, flourishing neighborhoods in Chicago and what distinguishes them from 'languishing', 'crime ridden' ones)?

This is a question of importance for multiple stakeholders:

- Individuals and agencies active in the field of crime prevention and/or responsible for decisions on deployment of police force.
- Individuals and agencies (incl. city government) active in the area of neighborhood social and business development, who are responsible to plan and budget such projects based on need and potential outcome.
- Individuals and business, who need to decide where to locate and settle themselves in the Chicago area.

Data requirements and -sources:

For the proposed study, two types of data will be required: 1. A dataset that represents business venue activity per neighborhood in Chicago 2. A dataset that contains an overview of crimes committed in the Chicago area incl. the locations of the crimes committed.

Dataset #1 can be constructed out of various data sources: a) An overview of Chicago Neighborhoods incl. geospatial definitions of the neighborhood boundaries for visualization and data exploration purposes. Datasets are available in multiple formats such as GeoJSON from [Ref3]. b) Data on venues incl. types of venues, longitude and latitude, with the flexibility to collect said data for each neighborhood within a perimeter from the center of the neighborhood. Such data is available via Foursquare [Ref4].

Dataset #2 is based on an annual overview of all crimes committed in the Chicago area, incl. type of crimes, longitude and latitude and so forth. Such a dataset is provided via the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system [Ref5].

Both datasets are processed, linked, and correlated to solve the questions posed in the previous section. The utilized analytical approach will be described in more detail in the next section.

Methodology:

The approach chosen for the data analysis follows 4 main steps:

1. Setting up and visualizing the geographical information of the Chicago neighborhoods.
2. Extracting the relevant crime statistics data, as well as linking to and aggregating with respect to the Chicago Neighborhood geographical information.
3. Retrieving and analyzing the relevant venue information based on the geographical characteristics of the Chicago neighborhoods.
4. Comparing and analyzing the communalities and differences between the crime and venue characteristics of the Chicago neighborhoods as extracted in steps 2 and 3, respectively.

1. Setting up and visualizing the geographical information of the Chicago neighborhoods

The main package used for visualization of geographical information is Folium [Ref6]. In order to zero in on the right portion of the world map, the center latitude and longitude of Chicago are retrieved via geopy's Nomatim function [Ref7].

A GeoJSON file of the geographical information of the Chicago neighborhoods was retrieved from [Ref3], see Tab.1. The boundaries of each Neighborhood (98 in total) contained in the GeoJSON file as polygon coordinates were then added to the Chicago map generated by Folium, using Folium's own built-in functionality. For visualization and further data mining purposes (such as done in Step 3), the center point as well as the "characteristic" radius of each neighborhood have to be determined. This was done by first converting the GeoJSON file into a geographical dataframe using the geopandas package [Ref8]. Then the centroids (in latitude and longitude), as well as the area (after projecting and rescaling of the coordinates in metric units) of each neighborhood were determined using geopandas' built-in functionality. The square root of the latter after division by pi yields the characteristic radius. The centroids were added as markers to the Folium map of Chicago, overlayed with the neighborhood boundaries, as shown in Fig. 1.

	pri_neigh	sec_neigh	shape_area	shape_len	geometry
0	Grand Boulevard	BRONZEVILLE	48492503.1554	28196.837157	MULTIPOLYGON (((-87.60671 41.81681, -87.60670 ...
1	Printers Row	PRINTERS ROW	2162137.97139	6864.247156	MULTIPOLYGON (((-87.62761 41.87437, -87.62760 ...
2	United Center	UNITED CENTER	32520512.7053	23101.363745	MULTIPOLYGON (((-87.66707 41.88885, -87.66707 ...
3	Sheffield & DePaul	SHEFFIELD & DEPAUL	10482592.2987	13227.049745	MULTIPOLYGON (((-87.65833 41.92166, -87.65835 ...
4	Humboldt Park	HUMBOLDT PARK	125010425.593	46126.751351	MULTIPOLYGON (((-87.74060 41.88782, -87.74060 ...
5	Garfield Park	GARFIELD PARK	89976069.5947	44460.91922	MULTIPOLYGON (((-87.69540 41.88819, -87.69520 ...
6	North Lawndale	NORTH LAWNDALE	89487422.0244	44959.459663	MULTIPOLYGON (((-87.72024 41.86987, -87.71965 ...
7	Little Village	LITTLE VILLAGE	127998297.819	49904.04003	MULTIPOLYGON (((-87.68740 41.83480, -87.68796 ...
8	Armour Square	ARMOUR SQUARE,CHINATOWN	17141468.6356	24359.189625	MULTIPOLYGON (((-87.62920 41.84713, -87.62919 ...
9	Avalon Park	AVALON PARK,CALUMET HEIGHTS	34852737.7366	27630.822534	MULTIPOLYGON (((-87.58566 41.75150, -87.58475 ...
10	Burnside	CHATHAM,BURNSIDE	16995983.2737	18137.944253	MULTIPOLYGON (((-87.58737 41.72326, -87.58733 ...
11	Hermosa	BELMONT CRAIGIN,HERMOSA	32602059.4055	27179.017438	MULTIPOLYGON (((-87.73146 41.93173, -87.73139 ...
12	Avondale	IRVING PARK,AVONDALE	55290595.482	34261.933404	MULTIPOLYGON (((-87.68799 41.93610, -87.68798 ...
13	Logan Square	LOGAN SQUARE	71256806.5518	36361.508518	MULTIPOLYGON (((-87.68732 41.91399, -87.68751 ...
14	Calumet Heights	AVALON PARK,CALUMET HEIGHTS	48826241.5643	32925.365871	MULTIPOLYGON (((-87.54568 41.72282, -87.54600 ...
15	East Side	SOUTHEAST SIDE	83241728.0493	52274.192849	MULTIPOLYGON (((-87.52462 41.69180, -87.52501 ...
16	West Pullman	WEST PULLMAN	99365198.0822	50023.843001	MULTIPOLYGON (((-87.61828 41.65911, -87.61829 ...
17	Garfield Ridge	MIDWAY AIRPORT	117890778.429	60080.44797	MULTIPOLYGON (((-87.73856 41.81871, -87.73853 ...
18	New City	BACK OF THE YARDS	134636963.254	48133.595961	MULTIPOLYGON (((-87.63546 41.79448, -87.63599 ...
19	Englewood	ENGLEWOOD	173600015.009	56144.046048	MULTIPOLYGON (((-87.62854 41.79414, -87.62853 ...

Table 1: First 20 rows of the dataframe of the Chicago neighborhoods extracted from [Ref3] using geopandas. It contains primary and secondary naming information (the secondary naming being discarded due to duplicates, and the primary naming being "semi-official" and unambiguous), as well as information on shape area and circumference. The information required for geographical operations as well as for visualizing the polygons representing each neighborhood is contained in the "geometry" column.

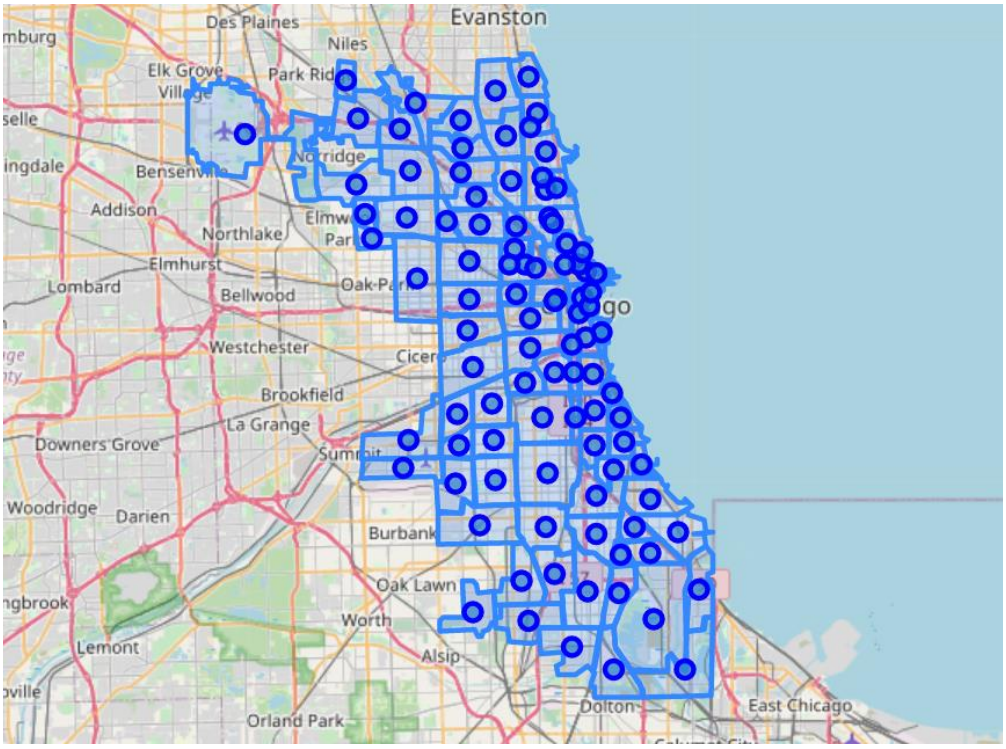

Figure 1: Map of the neighborhoods of Chicago with markers representing the geographic centers of each neighborhood. In the interactive version of the map, the name of each neighborhood will be displayed when clicking on a marker.

2. Extracting the relevant crime statistics data, as well as linking to and aggregating with respect to the Chicago Neighborhood geographical information.

A csv file containing all crimes registered in Chicago up until 1 year prior to the download date was retrieved from [Ref5] (containing around 200000 crimes) and read into a pandas dataframe [Ref9]. The relevant features contained in the dataframe next to the type (as represented by the primary and secondary description columns) are the latitude and longitude of the location of the occurrence of the crime. The dataframe was then converted into a geopandas [Ref8] dataframe , with the longitude and latitude columns being combined into a geometry object column for further processing in geopandas. This geometric representation of the crime location information was then used to map each crime to the neighborhood it occurred in by determining for each crime in the dataframe within which neighborhood polygon in Tab. 1 its location was falling. As a result, some rows with missing location information were dropped in the process. The resulting dataframe is shown in (Tab.2).

Subsequently, the number of crimes per neighborhood in Tab. 2 was counted and the result stored into a new dataframe (Tab.3). This dataframe was subsequently used to add a chloropleth map of the number of crimes per neighborhood to the map shown in Fig.1 using Folium. Further statistical analysis was then done to determine the top-10 neighborhoods with most and least crimes, respectively, and to determine and visualize the most frequently occurring types of crimes (as well as their relative frequency of occurrence) for both cases (Tab. 4 & Tab.5). Here, a new column 'Total Description' was generated by concatenating the primary and secondary description columns with a ": " delimiter, in order to generate a feature with maximized descriptive "granularity".

In order to further characterize the neighborhoods by the types of crimes occurring in them, a clustering analysis was executed. For this purpose, a new dataframe was generated, by first converting each crime in Tab.2 into a one-hot encoding per type of crime and then calculating the mean per resulting crime type column and neighborhood to yield the relative frequency of occurrence of the crime (Tab.6)

PRIMARY DESCRIPTION	SECONDARY DESCRIPTION	LATITUDE	LONGITUDE	pri_neigh
NARCOTICS	POSSESS - CRACK	41.819174	-87.628125	Grand Boulevard
DECEPTIVE PRACTICE	BOGUS CHECK	41.810546	-87.606588	Grand Boulevard
ROBBERY	STRONG ARM - NO WEAPON	41.805121	-87.621104	Grand Boulevard
CRIMINAL DAMAGE	TO VEHICLE	41.803121	-87.609460	Grand Boulevard
BATTERY	DOMESTIC BATTERY SIMPLE	41.808553	-87.622816	Grand Boulevard

Table 2: First 5 rows of the relevant columns of the dataframe of the Chicago crime statistics as extracted from [Ref5], with the last column containing the results of the mapping of the longitudinal and latitudinal location information to the neighborhoods contained in Tab.1.

	Neighborhood	CrimeCount
0	Austin	11997
1	Englewood	10930
2	Garfield Park	8616
3	South Shore	7656
4	Humboldt Park	7540
5	North Lawndale	7032
6	Auburn Gresham	6249
7	Roseland	5917
8	Grand Crossing	5754
9	Chatham	5180

Table 3: First 10 rows of the dataframe showing the total number of crimes per neighborhood.

Neighborhood	CrimeCount
Austin	11997
Englewood	10930
Garfield Park	8616
South Shore	7656
Humboldt Park	7540
North Lawndale	7032
Auburn Gresham	6249
Roseland	5917
Grand Crossing	5754
Chatham	5180

TOTAL DESCRIPTION	Percentage
BATTERY: DOMESTIC BATTERY SIMPLE	12.28
THEFT: $500 AND UNDER	6.98
CRIMINAL DAMAGE: TO PROPERTY	6.73
ASSAULT: SIMPLE	5.88
CRIMINAL DAMAGE: TO VEHICLE	5.22
BATTERY: SIMPLE	4.99
MOTOR VEHICLE THEFT: AUTOMOBILE	4.23
WEAPONS VIOLATION: UNLAWFUL POSSESSION - HANDGUN	3.49
THEFT: OVER $500	3.39
BURGLARY: FORCIBLE ENTRY	2.34

Table 4: Top 10 crime-ridden neighborhoods (left) and the 10 most frequent types of crimes occurring there (right).

Neighborhood	CrimeCount
Wrigleyville	313
Printers Row	285
Boystown	266
Andersonville	261
Jackson Park	212
Edison Park	209
Greektown	164
Grant Park	123
Millenium Park	122
Museum Campus	53

TOTAL DESCRIPTION	Percentage
THEFT: RETAIL THEFT	10.51
THEFT: $500 AND UNDER	9.31
THEFT: OVER $500	7.17
BATTERY: SIMPLE	7.12
THEFT: FROM BUILDING	5.08
MOTOR VEHICLE THEFT: AUTOMOBILE	4.63
BATTERY: DOMESTIC BATTERY SIMPLE	4.43
ASSAULT: SIMPLE	4.33
CRIMINAL DAMAGE: TO PROPERTY	4.03
CRIMINAL DAMAGE: TO VEHICLE	3.64

Table 5: Bottom 10 crime-ridden neighborhoods (left) and the 10 most frequent types of crimes occurring there (right).

	Neighborhood	ARSON: AGGRAVATED	ARSON: ATTEMPT ARSON	ARSON: BY EXPLOSIVE	ARSON: BY FIRE	ARSON: POSSESSION - EXPLOSIVE / INCENDIARY DEVICE	ASSAULT: AGG PO HANDS NO/MIN INJURY	ASSAULT: AGG PRO.EMP: HANDGUN	ASSAULT: AGG PRO.EMP: OTHER DANG WEAPON	ASSAULT: AGG PRO.EMP: OTHER FIREARM	...	WEAPONS VIOLATION: UNLAWFUL SALE - HANDGUN	WEAP(VIOLATI UNLAW S, HAND(
0	Albany Park	0.000000	0.000513	0.0	0.001026	0.0	0.000000	0.0	0.0	0.0	...	0.000000	
1	Andersonville	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	
2	Archer Heights	0.000000	0.000000	0.0	0.004418	0.0	0.000000	0.0	0.0	0.0	...	0.000000	
3	Armour Square	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	
4	Ashburn	0.000000	0.000000	0.0	0.003808	0.0	0.000544	0.0	0.0	0.0	...	0.000000	
...	...	...	...	...	...	...	...	...	...	...	...	...	
93	West Ridge	0.000000	0.000000	0.0	0.001373	0.0	0.000343	0.0	0.0	0.0	...	0.000000	
94	West Town	0.000000	0.000541	0.0	0.001081	0.0	0.000000	0.0	0.0	0.0	...	0.000000	
95	Wicker Park	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	
96	Woodlawn	0.000348	0.000000	0.0	0.001393	0.0	0.000697	0.0	0.0	0.0	...	0.000348	
97	Wrigleyville	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	

Table 6: Relative frequency of occurrence per neighborhood (rows) and crime type (columns).

The dataframe as shown in Tab.6 is then used as input for the clustering analysis, with each Neighborhood (aka each row in the data frame) being represented by a vector of relative occurrences of crimes as represented by the remaining columns of the dataframe. Clustering is done using the kMeans algorithm from the Scikit-learn package [Ref10]. In order to provide the appropriate number of clusters k as input for the clustering algorithm, clustering is first executed for a sequence of increasing k-Values, and for each k, the distortion (sum of squares of distance of each point to the center of its assigned cluster, [Ref11]) and average Silhouette over all points (a measure of 'clustering quality' based on the ratio of distances of each point to its own and nearest other cluster, [Ref12]) are calculated.

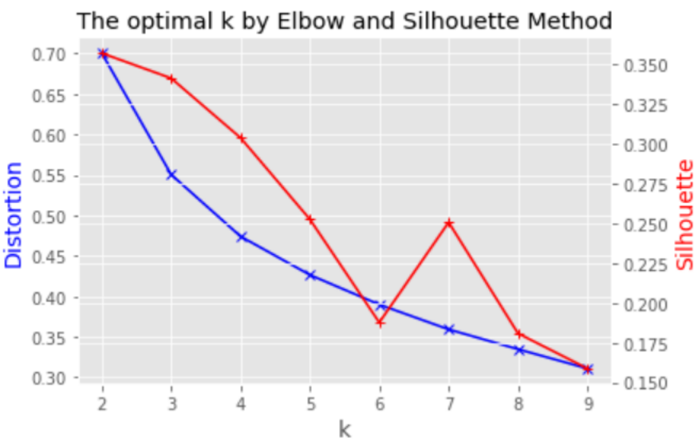

Figure 2: Determining the appropriate number of clusters using a combination of Elbow and Silhouette method for grouping of neighborhoods by occurrence of different types of crimes.

Then, the k-value for which the change in distortion starts to 'flatten' ("elbow method", [Ref11]) while maintaining a high average silhouette (meaning generally good clustering), is determined. Here, the number of groups k for the final cluster calculation is chosen to be 3. The final clustering result is visualized together with the crime density on top of the map of Chicago neighborhoods as shown in Fig.1. For the purpose of visualizing the assignment of each neighborhood to a cluster, the corresponding marker in the neighborhood center is colored accordingly.

To get more insights into the characteristics of each cluster, the dataframe in Tab.6 is further processed by sorting the columns per row and thereby determining the 10 most common crimes per neighborhood (Tab.7). When linked to the clustering result, the characteristic composition of the most frequent crimes per cluster can be determined.

	Neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime	6th Most Common Crime	7th Most Common Crime	8th Most Common Crime	9th Most Common Crime	10th Mos Commo Crim
0	Albany Park	THEFT: $500 AND UNDER	BATTERY: DOMESTIC BATTERY SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	BATTERY: SIMPLE	CRIMINAL DAMAGE: TO PROPERTY	THEFT: OVER $500	ASSAULT: SIMPLE	MOTOR VEHICLE THEFT: AUTOMOBILE	THEFT: RETAIL THEFT	BURGLAR FORCIBL ENTR
1	Andersonville	THEFT: RETAIL THEFT	THEFT: $500 AND UNDER	THEFT: OVER $500	CRIMINAL DAMAGE: TO PROPERTY	MOTOR VEHICLE THEFT: AUTOMOBILE	ASSAULT: SIMPLE	BURGLARY: UNLAWFUL ENTRY	DECEPTIVE PRACTICE: FINANCIAL IDENTITY THEFT $...	THEFT: FROM BUILDING	BATTER SIMPL
2	Archer Heights	BATTERY: DOMESTIC BATTERY SIMPLE	THEFT: $500 AND UNDER	THEFT: RETAIL THEFT	THEFT: OVER $500	CRIMINAL DAMAGE: TO PROPERTY	ASSAULT: SIMPLE	MOTOR VEHICLE THEFT: AUTOMOBILE	BATTERY: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	WEAPON VIOLATION RECKLES FIREAR DISCHARG
3	Armour Square	BATTERY: DOMESTIC BATTERY SIMPLE	BATTERY: SIMPLE	THEFT: $500 AND UNDER	ASSAULT: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	THEFT: OVER $500	MOTOR VEHICLE THEFT: AUTOMOBILE	CRIMINAL DAMAGE: TO PROPERTY	WEAPONS VIOLATION: UNLAWFUL POSSESSION - HANDGUN	BURGLAR FORCIBL ENTR
4	Ashburn	BATTERY: DOMESTIC BATTERY SIMPLE	THEFT: $500 AND UNDER	CRIMINAL DAMAGE: TO PROPERTY	THEFT: OVER $500	ASSAULT: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	MOTOR VEHICLE THEFT: AUTOMOBILE	BATTERY: SIMPLE	DECEPTIVE PRACTICE: FINANCIAL IDENTITY THEFT $...	BURGLAR FORCIBL ENTR

Table 7: 10 most common types of crime per neighborhood, first 5 rows.

3. Retrieving and analyzing the relevant venue information based on the geographical characteristics of the Chicago neighborhoods.

The Foursquare API [Ref4] was used to retrieve the top 100 venues per neighborhood based on each neighborhood's geographical information contained in the dataframe as shown in Tab. 1. As input for the API query, the latitude and longitude of the center point, as well the characteristic radius of each neighborhood, as calculated in Step 1, were used, in order to tailor the search reference location and radius to each individual neighborhood. Especially the latter is necessary since it is obvious from Fig.1. that neighborhoods vary considerably in size. Even when making the search adaptive as described, due to the irregular geometries of the neighborhoods, venues might be falsely assigned to a certain neighborhood. In order to check and if required correct the assignment of a venue to a neighborhood, the venue longitude and latitude as returned by the Foursquare API are mapped in a similar fashion to the neighborhood geographical information, as described in Step 2 for the crime information. The resulting dataframe is shown in Tab. 8.

Since the data obtained in the way described above is likely not offering a complete representation of the business venue activity (e.g. due to obvious biases in the user based reporting/information gathering of Foursquare and the number of venues retrieved), conclusions on the density of business activity cannot easily be drawn. However, the characteristics of business venue activity may be compared between neighborhoods by utilizing a similar grouping approach as described in Step 2 to group neighborhoods according to the types of venues based on the dataframe shown in Tab.8. The resulting data frame with the relative frequencies of occurrence of each type of venue used as grouping input is shown in Tab.9. Again, a combination of elbow and silhouette method was utilized to determine the appropriate number of groups in which the data is grouped (Fig.3). To facilitate the comparison of the groupings obtained for the crime and venue data, respectively, the same k value as in Step 2 is utilized, which should be not too far off from the desired value. The data in tab.9 is then again reshaped into the representation shown in tab.10, in order to compare the characteristics of the different clusters in terms of composition of venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	geometry
0	Grand Boulevard	41.812949	-87.617860	Chicago Blues District	41.810071	-87.614105	Jazz Club	POINT (-87.61411 41.81007)
1	Grand Boulevard	41.812949	-87.617860	Blues Brothers Mural / Shelly's Loan & Jewelry...	41.809391	-87.619517	Plaza	POINT (-87.61952 41.80939)
2	Grand Boulevard	41.812949	-87.617860	Peach's Restaurant	41.809481	-87.617009	Breakfast Spot	POINT (-87.61701 41.80948)
3	Grand Boulevard	41.812949	-87.617860	Ain't She Sweet Cafe	41.816817	-87.613004	Coffee Shop	POINT (-87.61300 41.81682)
4	Grand Boulevard	41.812949	-87.617860	Parkway Ballroom	41.813142	-87.616064	Food	POINT (-87.61606 41.81314)

Table 8: First 5 rows of the dataframe containing the top venues per neighborhood (incl. venue type and location).

	Neighborhood	Yoga Studio	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	...	Warehouse	Warehouse Store	Waterfront	Wei L Cei
0	Albany Park	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000
1	Andersonville	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000
2	Archer Heights	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000
3	Armour Square	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000
4	Ashburn	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.020

Table 9: Relative frequency of occurrence per neighborhood (rows) and venue type (columns), first 5 rows.

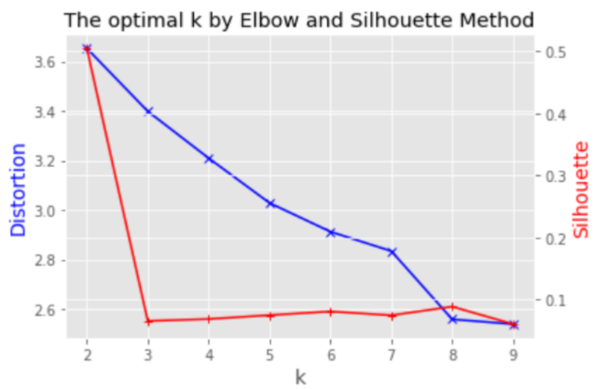

Figure 3: Determining the appropriate number of clusters using a combination of Elbow and Silhouette method for grouping of neighborhoods by occurrence of venues.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park	Chinese Restaurant	Ice Cream Shop	Thai Restaurant	Thrift / Vintage Store	Middle Eastern Restaurant	Bakery	Grocery Store	Asian Restaurant	Park	Sandwich Place
1	Andersonville	Breakfast Spot	Coffee Shop	Miscellaneous Shop	Grocery Store	Café	Bakery	Italian Restaurant	Pharmacy	New American Restaurant	Burger Joint
2	Archer Heights	Mexican Restaurant	Mobile Phone Shop	Bank	Pizza Place	Bar	Discount Store	Fast Food Restaurant	Seafood Restaurant	Sandwich Place	Pet Store
3	Armour Square	Baseball Stadium	Bar	Historic Site	Park	Bus Station	Lounge	Train Station	American Restaurant	Cosmetics Shop	Chinese Restaurant
4	Ashburn	Fast Food Restaurant	Park	Sandwich Place	Pizza Place	Fried Chicken Joint	Donut Shop	Liquor Store	Ice Cream Shop	American Restaurant	Convenience Store

Table 10: 10 most common types of venues per neighborhood, first 5 rows.

4. Comparing and analyzing the communalities and differences between the crime and venue characteristics of the Chicago neighborhoods as extracted in steps 2 and 3, respectively.

The clusters obtained in Step 2 and 3, respectively, are then compared for communality (i.e. overlap). This comparison is done by 2 means: adjusted rand index (ARI, [Ref13]) and confusion matrix [Ref14]. The ARI is a single, normalized, key performance indicator that indicates the degree of communality between 2 different cluster labelings on the same dataset. It can reach values between 0 (indicating no communality) and 1 (indicating perfect communality), independent of the actual cluster labelling (meaning the order of respective cluster labels in each dataset). The confusion matrix is then used to visualize to which degrees the individual clusters obtained in step 2 and 3, respectively, overlap, which may provide potential insights into underlying connections and shared causalities of crime and venue related characteristics of the different neighborhoods.

Results:

In this chapter, the results of the study are discussed in 3 sections:

1. Crime characteristics of Chicago neighborhoods
2. Characteristic venues of Chicago neighborhoods
3. Relationship between crime and venue characteristics of Chicago neighborhoods

1. Crime characteristics of Chicago neighborhoods

As shown already in Tab. 4 and 5, there is quite a difference in number of crimes in Chicago neighborhoods. The most crime ridden neighborhoods are Austin, Englewood and Garfield Park (combined East and West Garfield Park). The connecting main socio-demographics of these neighborhoods are the low median income (as low as 21000$) and the high poverty rate (up to 44%), [Ref15-17] . Combined with the fact that the vast majority of inhabitants is black, the high crime rate unfortunately appears to be strongly linked to the precarious social situation of the African American community living there. On the contrary, the lowest ranking neighborhoods in terms of number of crimes consist mostly of commercial, public and entertainment districts such as Greektown, Grant Park, Millenium Park and Museum Campus with a generally low percentage of permanent residents, [Ref18-21].
As shown in Figs. 4 and 5 (which visualize the content of Tab. 4 and 5), the most and least crime ridden neighborhoods also differ in types of crimes committed. Whereas violent crimes such as domestic violence, assault and criminal damage dominate the top-10 crime-ridden neighborhoods, the crimes most frequently committed in the bottom-10 neighborhoods mainly consist of various types of theft (with retail theft taking the top spot).
This observation is further solidified by the cluster analysis executed to distinguish neighborhoods based on the types of crimes committed in these. Fig. 6 shows the map of Chicago with the superimposed neighborhoods. The neighborhoods are colored according to the number of crimes occurring there over the course of a year, whereas the color of the center marker indicates its assignment to a certain cluster. Tabs. 11, 12 and 13 show the top-10 crimes in the top-5 neighborhoods of each cluster. In these overviews, the tables have been sorted according to the ‘Silhouette’ value of each neighborhood in descending order, meaning the most representative neighborhoods for each cluster appear at the top of the table.

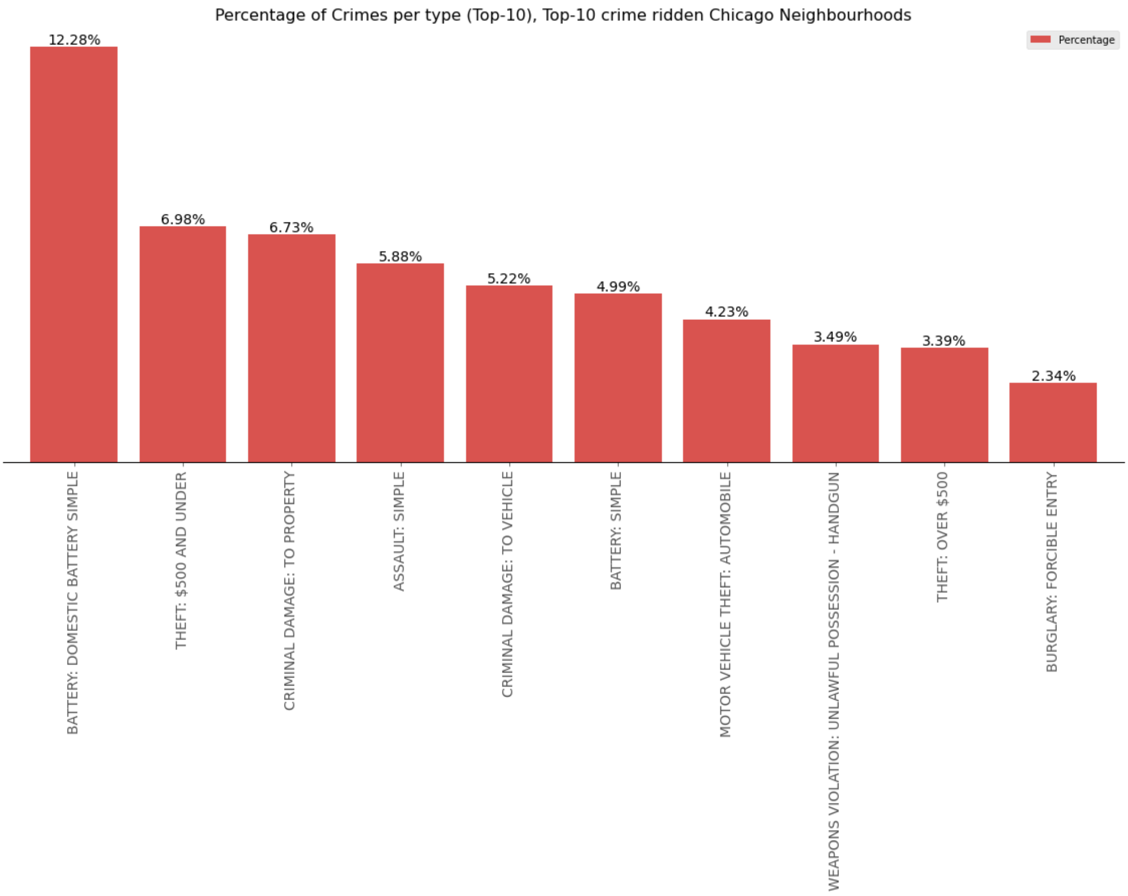

Figure 4: Bar chart of the relative frequency of the most common types of crimes in the top-10, most crime-ridden neighborhoods.

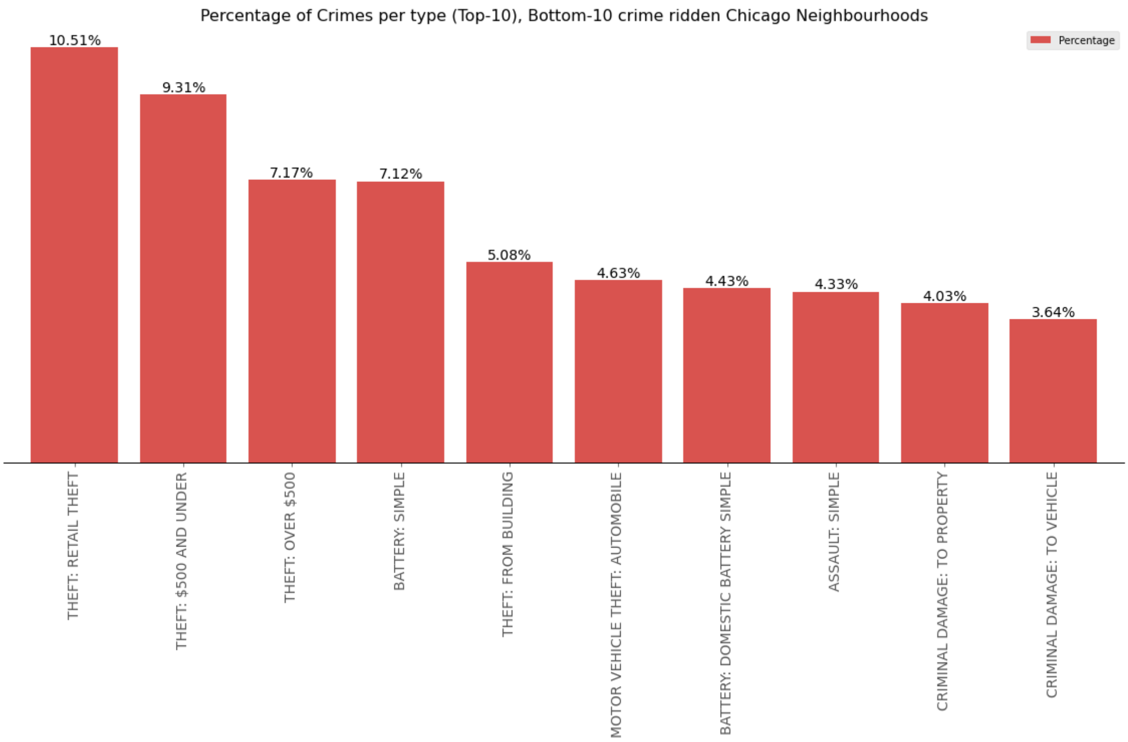

Figure 5: Bar chart of the relative frequency of the most common types of crimes in the bottom-10, least crime-ridden neighborhoods.

The first and most numerous cluster 0 (red dots in Fig. 6) is characterized by violent crimes such as (domestic) battery, criminal damage and assault, with the neighborhoods predominantly located in the West and South of Chicago, and in general showing elevated crime rates (Tab.11). The second most numerous cluster 1 (purple dots in Fig. 6), is mainly showing neighborhoods located to the North-East of Chicago, which are dominated by various types of theft (ranging from petty to motorvehicle and identity theft), mixed in with some violent crimes like battery and burglary (Tab. 12). The 3rd cluster 2 (green dots in Fig. 6, overlapped by cluster 1, located at the very East of Chicago) only consist of 2 neighborhoods, Greektown and Magnificent mile, and is mainly characterized by retail theft as the predominant type of crime (Tab. 13). This might be explained by the fact that both neighborhoods are renowned for their shops, entertainment businesses and restaurants [Ref18 & 22]. This cluster appears to be more similar to the second cluster 1 than the first cluster 0 in its characteristics.

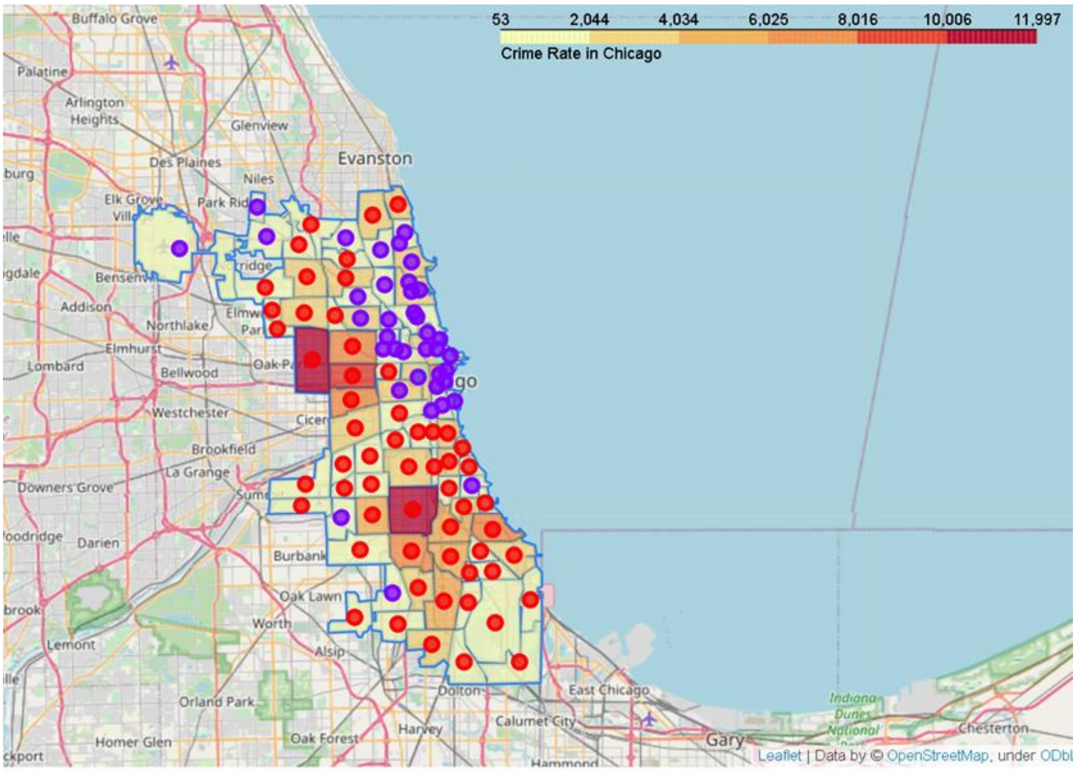

Figure 6: Chloropleth map of the Chicago neighborhoods, where the shading of each neighborhood indicates the number of registered crimes over 1 year and the color of each "centroid" its cluster assignment with respect to types of crimes committed.

pri_neigh	Silhouette	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime	6th Most Common Crime	7th Most Common Crime	8th Most Common Crime	9th Most Common Crime	10th Most Common Crime
Chicago Lawn	0.609052	BATTERY: DOMESTIC BATTERY SIMPLE	THEFT: $500 AND UNDER	CRIMINAL DAMAGE: TO PROPERTY	ASSAULT: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	BATTERY: SIMPLE	THEFT: OVER $500	MOTOR VEHICLE THEFT: AUTOMOBILE	BURGLARY: FORCIBLE ENTRY	WEAPONS VIOLATION: RECKLESS FIREARM DISCHARGE
South Shore	0.597854	BATTERY: DOMESTIC BATTERY SIMPLE	CRIMINAL DAMAGE: TO PROPERTY	THEFT: $500 AND UNDER	ASSAULT: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	BATTERY: SIMPLE	MOTOR VEHICLE THEFT: AUTOMOBILE	THEFT: OVER $500	BURGLARY: FORCIBLE ENTRY	OTHER OFFENSE: TELEPHONE THREAT
Grand Crossing	0.596157	BATTERY: DOMESTIC BATTERY SIMPLE	CRIMINAL DAMAGE: TO PROPERTY	THEFT: $500 AND UNDER	ASSAULT: SIMPLE	MOTOR VEHICLE THEFT: AUTOMOBILE	CRIMINAL DAMAGE: TO VEHICLE	BATTERY: SIMPLE	WEAPONS VIOLATION: UNLAWFUL POSSESSION - HANDGUN	THEFT: OVER $500	BURGLARY: FORCIBLE ENTRY
Auburn Gresham	0.595385	BATTERY: DOMESTIC BATTERY SIMPLE	THEFT: $500 AND UNDER	CRIMINAL DAMAGE: TO PROPERTY	ASSAULT: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	BATTERY: SIMPLE	MOTOR VEHICLE THEFT: AUTOMOBILE	THEFT: OVER $500	WEAPONS VIOLATION: UNLAWFUL POSSESSION - HANDGUN	BURGLARY: FORCIBLE ENTRY
Woodlawn	0.592744	BATTERY: DOMESTIC BATTERY SIMPLE	THEFT: $500 AND UNDER	CRIMINAL DAMAGE: TO PROPERTY	BATTERY: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	ASSAULT: SIMPLE	THEFT: OVER $500	MOTOR VEHICLE THEFT: AUTOMOBILE	WEAPONS VIOLATION: UNLAWFUL POSSESSION - HANDGUN	BURGLARY: FORCIBLE ENTRY

Table 11: Top-10 crimes of the 5 most representative neighborhoods for cluster 0.

pri_neigh	Silhouette	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime	6th Most Common Crime	7th Most Common Crime	8th Most Common Crime	9th Most Common Crime	10th Most Common Crime
Lake View	0.360684	THEFT: $500 AND UNDER	THEFT: RETAIL THEFT	THEFT: OVER $500	THEFT: FROM BUILDING	BATTERY: SIMPLE	MOTOR VEHICLE THEFT: AUTOMOBILE	ASSAULT: SIMPLE	CRIMINAL DAMAGE: TO PROPERTY	DECEPTIVE PRACTICE: FINANCIAL IDENTITY THEFT $...	CRIMINAL DAMAGE: TO VEHICLE
Lincoln Park	0.349276	THEFT: $500 AND UNDER	THEFT: OVER $500	THEFT: RETAIL THEFT	THEFT: FROM BUILDING	BATTERY: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	CRIMINAL DAMAGE: TO PROPERTY	ASSAULT: SIMPLE	BATTERY: DOMESTIC BATTERY SIMPLE	BURGLARY: FORCIBLE ENTRY
Wicker Park	0.330413	THEFT: OVER $500	THEFT: RETAIL THEFT	THEFT: $500 AND UNDER	BURGLARY: FORCIBLE ENTRY	THEFT: FROM BUILDING	BATTERY: SIMPLE	ASSAULT: SIMPLE	CRIMINAL DAMAGE: TO PROPERTY	MOTOR VEHICLE THEFT: AUTOMOBILE	BATTERY: DOMESTIC BATTERY SIMPLE
Bucktown	0.327667	THEFT: $500 AND UNDER	THEFT: OVER $500	THEFT: RETAIL THEFT	MOTOR VEHICLE THEFT: AUTOMOBILE	THEFT: FROM BUILDING	BATTERY: SIMPLE	BATTERY: DOMESTIC BATTERY SIMPLE	DECEPTIVE PRACTICE: FINANCIAL IDENTITY THEFT $...	CRIMINAL DAMAGE: TO VEHICLE	CRIMINAL DAMAGE: TO PROPERTY
West Loop	0.305438	THEFT: OVER $500	THEFT: RETAIL THEFT	THEFT: $500 AND UNDER	MOTOR VEHICLE THEFT: AUTOMOBILE	THEFT: FROM BUILDING	BATTERY: SIMPLE	ASSAULT: SIMPLE	CRIMINAL DAMAGE: TO VEHICLE	BURGLARY: FORCIBLE ENTRY	BATTERY: DOMESTIC BATTERY SIMPLE

Table 12: Top-10 crimes of the 5 most representative neighborhoods for cluster 1.

pri_neigh	Silhouette	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime	6th Most Common Crime	7th Most Common Crime	8th Most Common Crime	9th Most Common Crime	10th Most Common Crime
Magnificent Mile	0.636912	THEFT: RETAIL THEFT	BURGLARY: FORCIBLE ENTRY	THEFT: FROM BUILDING	DECEPTIVE PRACTICE: CREDIT CARD FRAUD	THEFT: OVER $500	THEFT: $500 AND UNDER	THEFT: POCKET-PICKING	BATTERY: SIMPLE	CRIMINAL DAMAGE: TO PROPERTY	ASSAULT: SIMPLE
Greektown	0.563680	THEFT: RETAIL THEFT	THEFT: $500 AND UNDER	BATTERY: SIMPLE	THEFT: OVER $500	ASSAULT: SIMPLE	THEFT: FROM BUILDING	CRIMINAL DAMAGE: TO PROPERTY	MOTOR VEHICLE THEFT: AUTOMOBILE	BURGLARY: FORCIBLE ENTRY	BATTERY: DOMESTIC BATTERY SIMPLE

Table 13: Top-10 crimes of the 2 neighborhoods contained in cluster 2.

2. Characteristic venues of Chicago neighborhoods

A similar cluster analysis has been performed on venue data retrieved via the Foursquare API and mapped to the Chicago neighborhoods. The result is visualized in Fig. 7. Here, for better visual comparison, the same Chloropleth map representing the number of crimes per neighborhood as in Fig. 6 is overlapped with neighborhood centroids that have been colored according to their cluster assignment. KMeans clustering with a k of 3 yields 2 large main and 1 smaller cluster. The first main cluster 0 (red dots in Fig. 7) is mainly localized around the North-East of Chicago, similar to cluster 1 for the crime-based analysis. It mainly consist of neighborhoods whose top-venues are bars, restaurants and coffeeshops (Tab. 14). Cluster 1 (purple dots in Fig. 7) only contains 3 neighborhoods, mainly characterized by parks and outside activities (Tab. 15). Cluster 2 (green dots in Fig. 7) is mainly characterized by low cost fast food joints and discount stores (Tab. 15), and is located in the West and South of Chicago, geographically coinciding with cluster 0 from the crime-based analysis.

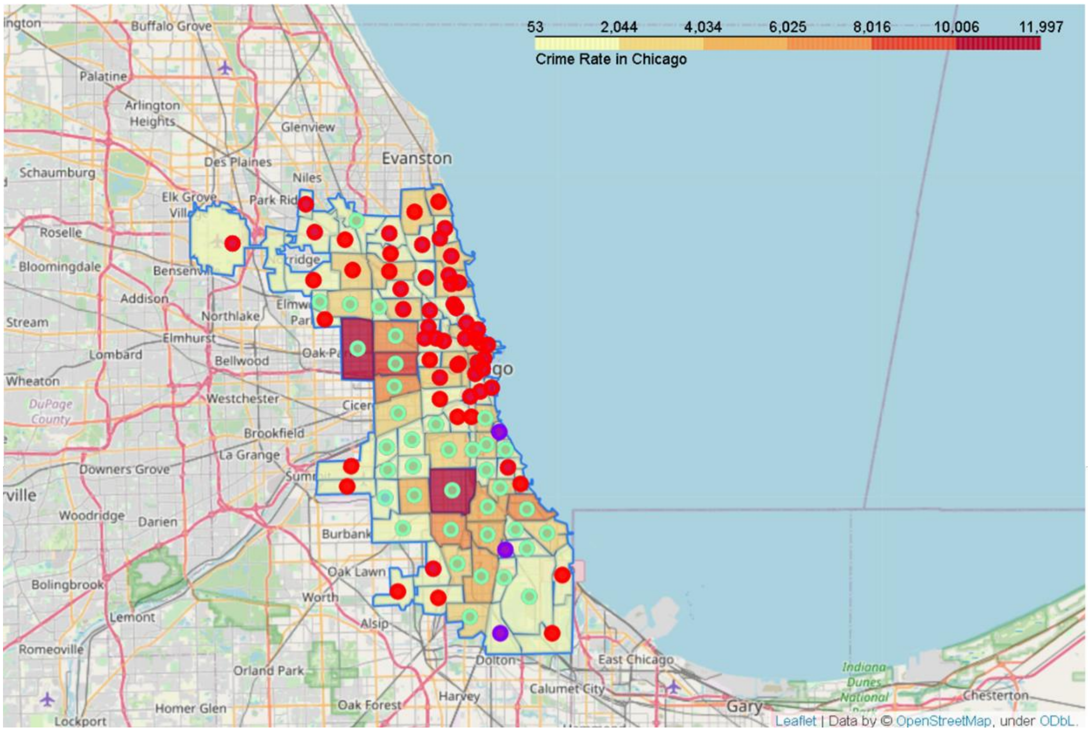

Figure 7: Chloropleth map of the Chicago neighborhoods, where the shading of each neighborhood indicates the number of registered crimes over 1 year and the color of each "centroid" its cluster assignment with respect to types of top venues present in the neighborhood.

pri_neigh	Silhouette	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
West Loop	0.112451	Coffee Shop	Italian Restaurant	Café	Dance Studio	Deli / Bodega	Grocery Store	Pizza Place	Bar	New American Restaurant	Gym / Fitness Center
Wicker Park	0.109663	Pizza Place	Bakery	Italian Restaurant	Bar	Coffee Shop	Sushi Restaurant	Bookstore	Nail Salon	Salon / Barbershop	Gourmet Shop
Bucktown	0.104620	Bar	Coffee Shop	Hot Dog Joint	French Restaurant	Gym	Korean Restaurant	Grocery Store	Mexican Restaurant	Park	Pizza Place
Old Town	0.090020	Pizza Place	Gym / Fitness Center	Bar	Comedy Club	Gym	Italian Restaurant	Burger Joint	Cycle Studio	Molecular Gastronomy Restaurant	Coffee Shop
North Center	0.089006	Bar	Coffee Shop	Brewery	Mexican Restaurant	Music Venue	Antique Shop	Sandwich Place	Park	Sushi Restaurant	Grocery Store

Table 14: Top-10 crimes of the top-5 neighborhoods contained in cluster 0.

pri_neigh	Silhouette	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Burnside	0.175612	Park	Intersection	Train Station	North Indian Restaurant	Pakistani Restaurant	Outlet Store	Outdoors & Recreation	Outdoor Sculpture	Other Repair Shop	Other Great Outdoors
Riverdale	-0.041991	Park	Harbor / Marina	Residential Building (Apartment / Condo)	Garden	Grocery Store	North Indian Restaurant	Outlet Store	Outdoors & Recreation	Outdoor Sculpture	Other Repair Shop
Oakland	-0.117170	Beach	Park	Track	Hotel	Vineyard	Juice Bar	Dog Run	Athletics & Sports	Climbing Gym	Public Art

Table 15: Top-10 crimes of the 3 neighborhoods contained in cluster 1. The negative silhouette values of the Riverdale and Oakland neighborhoods suggest that their cluster assignment may be questionable at best.

pri_neigh	Silhouette	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Austin	0.238561	Fast Food Restaurant	Discount Store	Park	Sandwich Place	Fried Chicken Joint	Grocery Store	Donut Shop	Pharmacy	Mexican Restaurant	Clothing Store
Roseland	0.217754	Fast Food Restaurant	Fried Chicken Joint	Sandwich Place	Donut Shop	Discount Store	Liquor Store	Park	Shop & Service	Grocery Store	Food
Englewood	0.213879	Fast Food Restaurant	Discount Store	Sandwich Place	Donut Shop	Gas Station	Grocery Store	Clothing Store	Fried Chicken Joint	Bank	Shoe Store
Chicago Lawn	0.193539	Mexican Restaurant	Discount Store	Fast Food Restaurant	Donut Shop	Bakery	Sandwich Place	Pizza Place	American Restaurant	Currency Exchange	Food
Auburn Gresham	0.191861	Fast Food Restaurant	Discount Store	Pharmacy	Seafood Restaurant	Lounge	Fried Chicken Joint	Sandwich Place	Southern / Soul Food Restaurant	Chinese Restaurant	Grocery Store

Table 16: Top-10 crimes of the top-5 neighborhoods contained in cluster 2.

3. Relationship between crime and venue characteristics of Chicago neighborhoods

A comparison of the crime- and venue based clusters by means of the ARI yields a value of **0.27**. That is higher than 0 (result similar to "random" clustering) but not particularly high. However, it could be affected by the differing size of the clusters. As an alternative method, the Jaccard Similarity Index [Ref23] was calculated both globally (counting the total true positives, false negatives and false positives) as well as based on the size-weighted average of the score per individual cluster label. Here, one set of clusters had to be relabeled according to the best matching clusters in the other set. In both cases, values close to **0.60** were found, indicating fairly good similarity between the clusters.

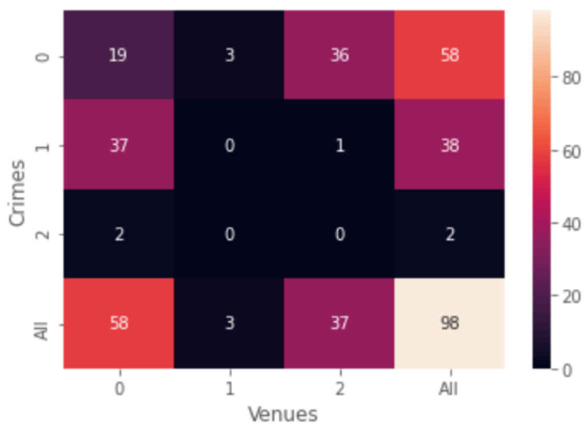

Figure 8: Confusion matrix showing the number of matching observations between crime and venue-based clusters.

To visualize the correspondence between crime and venue-based clusters, a crosstable was calculated and visualized as a confusion matrix, showing the number of matching neighborhoods between 2 clusters of each type, Fig. 8. Here, the observation from section 2 of this chapter can be confirmed, namely, that crime-based clusters 0 and 1 show the best match with venue based clusters 0 and 2, respectively.

Discussion:

From the results discussed in the previous chapter, some relationships between frequency and types of crimes committed and business venue activity can be concluded. More prosperous neighborhoods such as located more towards the North-East of Chicago exhibit a vast amount of high-profile business venue activity such as coffee shops, restaurants, bars, etc. Crime rate is generally lower in these neighborhoods, with the committed crimes being less severe (petty crimes like various types of theft). On the contrary, the less prosperous neighborhoods more located in the East and South-East of Chicago are dominated by less glamorous venues such as Fastfood restaurants and Discount stores. Crime rate is the highest in Chicago (and by extension in the US) in these areas, with the majority of crimes of more violent nature such as domestic battery and assault.

The likely underlying rootcause can be found in the socio-demographic characteristics of these neighborhoods. The majority of the population in the crime-ridden neighborhoods like Austin, Englewood and Garfield park lives in precarious circumstances with limited financial means. This might in turn make these neighborhoods less attractive for more high-profile business ventures due to a lack of matching “clientele”. In turn the resulting lack of employment and perspective may cause the high amount of (domestic) violence, which may pose another deterring factor for the settlement of business ventures in these areas. On the contrary, the more prosperous neighborhoods are either mainly offering high profile housing like lofts [Ref24] or are in-fact non-residential business and/or entertainment areas [Ref20-22]. These may in general attract more wealthy “clientele”, which appears to be reflected by the general overall smaller crime rate and overall more petty types of crimes committed.

It is not so straightforward to offer recommendations, let alone solutions to the problem. Establishing similar types of high-profile venues in the poorer areas will likely be doomed to fail due to lack of financial perspective and appropriate customer base. Even if successful, the problem will likely just be “moved” to other areas of Chicago, as real estate would become more expensive and less affordable to the current population as a result of the establishment of high-profile business venues. That is to be avoided, if one aspires to solve the issue of crime rate and poverty in a sustainable fashion for the whole of Chicago. One suggestion to make instead is to tackle the apparent issue of lack of diversity in the types of venues. Chicago seems to be mainly separated into 2 types of areas with their own socio-economic characteristics. Social-demographic development projects in the poor areas may help to create and attract businesses and venues “appropriate” to the neighborhoods, which may in turn help to generate employment possibilities as well as steadily increase the standard of living in these areas without the danger of forcing the original local population out of these neighborhoods.

Conclusion and Outlook:

In this investigation, I investigated the relationship between amount and types of crimes committed in the neighborhoods of Chicago and the characteristics of business venue activity found there. By means of a combination of geographical and cluster analysis I was able to identify which characteristics separates poor, crime-ridden neighborhoods from flourishing, ‘safe’ neighborhoods. These insights may be valuable for members and agencies of law enforcement to help them guide decisions concerning police presence, as well as for business developers and entrepreneurs, who have to decide where to locate which type of business venue. Based on my investigation I was also able to provide some general suggestions on how (not) to try to improve the situation in the poorer, more crime-ridden areas in a sustainable, long-term way, which may be valuable for individuals and agencies (incl. city government) in charge of socio-economic development programs.

However, it is to be noted here that more in-depth analysis will very likely yield more fundamental insights and more specific suggestions and recommendations than I was able to provide here. More data sources should be included and more aspects investigated in any future analysis on the subject matter, including for instances population demographics, educational situation, other types of business activity (e.g. industrial activity) and so forth. Also, in order to increase the correlation between the different feature sets, more elaborate data science and machine learning techniques may be employed that aim at maximizing co-variance through feature projection such as principal component analysis [Ref25].

Literature references:

[Ref1]: http://www.city-data.com/crime/crime-Chicago-Illinois.html, accessed 13-dec 2020.

[Ref2]: “What's Causing Chicago’s Homicide Spike?” by Matt Ford, The Atlantic, January 25, 2017, https://www.theatlantic.com/politics/archive/2017/01/chicago-homicide-spike-2016/514331/ , accessed 13-dec 2020.

[Ref3]: Neighborhood boundaries in Chicago, developed by the Office of Tourism: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/bbvz-uum9 , accessed 13-dec 2020.

[Ref4]: Foursquare website: https://foursquare.com/ , accessed 13-dec 2020.

[Ref5]: https://www.chicago.gov/city/en/dataset/crime.html , accessed 13-dec 2020.

[Ref6]: Folium: https://pypi.org/project/folium/ , accessed 2-jan 2021.

[Ref7]: Nomatim: https://nominatim.org/ , accessed 2-jan 2021.

[Ref8]: Geopandas: https://geopandas.org/ , accessed 2-jan 2021.

[Ref9]: Pandas: https://pandas.pydata.org/ , accessed 2-jan 2021.

[Ref10]: Scikit learn: https://scikit-learn.org/stable/index.html , accessed 2-jan 2021.

[Ref11]: “Who belongs in the family?” by Robert L. Thorndike , Psychometrika 18, 267–276 (1953).

[Ref12]: "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis" by Peter J. Rousseeuw, Computational and Applied Mathematics. 20: 53–65 (1987).

[Ref13]: "Objective criteria for the evaluation of clustering methods" by William M. Rand, Journal of the American Statistical Association 66 (336): 846–850 (1971).

[Ref14]: “Selecting and interpreting measures of thematic classification accuracy" by Stephen V. Stehman, Remote Sensing of Environment. 62 (1): 77–89 (1997).

[Ref15]: https://en.wikipedia.org/wiki/Austin,_Chicago , accessed 3-jan 2021.

[Ref16]: https://en.wikipedia.org/wiki/Englewood,_Chicago , accessed 3-jan 2021.

[Ref17]: https://en.wikipedia.org/wiki/West_Garfield_Park,_Chicago & https://en.wikipedia.org/wiki/East_Garfield_Park,_Chicago , accessed 3-jan 2021.

[Ref18]: https://en.wikipedia.org/wiki/Greektown,_Detroit , accessed 3-jan 2021.

[Ref19]: https://en.wikipedia.org/wiki/Grant_Park_(Chicago) , accessed 3-jan 2021.

[Ref20]: https://en.wikipedia.org/wiki/Millennium_Park , accessed 3-jan 2021.

[Ref21]: https://en.wikipedia.org/wiki/Museum_Campus , accessed 3-jan 2021.

[Ref22]: https://en.wikipedia.org/wiki/Magnificent_Mile , accessed 3-jan 2021.

[Ref23]: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html , accessed 3-jan 2021.

[Ref24]: https://en.wikipedia.org/wiki/Printer%27s_Row,_Chicago , accessed 3-jan 2021.

[Ref25]: "Analysis of a complex of statistical variables into principal components" by Harold Hotelling, Journal of Educational Psychology, 24, 417–441, and 498–520, (1933).